Introduction to Machine Learning

Preprocessig in Natural Language Processing

Andres Mendez-Vazquez

March 14, 2019

Outline

Outline

As Always, we need to Preprocess

For example, simply counting words can give you interesting info

- This is known as unigram word count (or word frequency, when normalized).

How do we count words?

- First we need to define what a word is.

This is highly non-trivial for languages without space

- For example Chinese...

As Always, we need to Preprocess

For example, simply counting words can give you interesting info

- This is known as unigram word count (or word frequency, when normalized).

How do we count words?

- First we need to define what a word is.

This is highly non-trivial for languages without space

- For example Chinese...

As Always, we need to Preprocess

For example, simply counting words can give you interesting info

- This is known as unigram word count (or word frequency, when normalized).

How do we count words?

- First we need to define what a word is.

This is highly non-trivial for languages without space

- For example Chinese...

Thus

The importance of defining the concept of a word

- For example Is a number a Word?
  - 124366
  - One hundred thousand

In another example, French

- L'ensemble one word or two?
- Google until 2008 could not
  - l'ensemble to match with un ensemble

With German is worse

- Lebensversicherungsgesellschaftsangestellter
- 'life insurance company employee'

Thus

The importance of defining the concept of a word

- For example Is a number a Word?
  - 124366
  - One hundred thousand

In another example, French

- L'ensemble one word or two?
- Google until 2008 could not
  - l'ensemble to match with un ensemble

With German is worse

- Lebensversicherungsgesellschaftsangestellter
- 'life insurance company employee'

Thus

The importance of defining the concept of a word

- For example Is a number a Word?
  - 124366
  - One hundred thousand

In another example, French

- L'ensemble one word or two?
- Google until 2008 could not
  - l'ensemble to match with un ensemble

With German is worse

- Lebensversicherungsgesellschaftsangestellter
- 'life insurance company employee'

Outline

Prepossessing text is called tokenization or text normalization

Sometimes you need to throw away stuff

- e.g., HTML tags – but sometimes they are valuable, UUencoding, etc.

Word boundaries

- White space and punctuations
  - But words like Ph.D., isn't, e-mail, C|net or $19.99 are problematic.

Actually at the end many people use regular expressions for this

- It can take years to build them...

Prepossessing text is called tokenization or text normalization

Sometimes you need to throw away stuff

- e.g., HTML tags – but sometimes they are valuable, UUencoding, etc.

Word boundaries

- White space and punctuations
  - But words like Ph.D., isn't, e-mail, C|net or $19.99 are problematic.

Actually at the end many people use regular expressions for this

- It can take years to build them...

Prepossessing text is called tokenization or text normalization

Sometimes you need to throw away stuff

- e.g., HTML tags – but sometimes they are valuable, UUencoding, etc.

Word boundaries

- White space and punctuations
  - But words like Ph.D., isn't, e-mail, C|net or $19.99 are problematic.

Actually at the end many people use regular expressions for this

- It can take years to build them...

Outline

Stop Words

Things with little value

- Common words which would appear to be of little value.
  - e.g. the, a, and, to, be

What is the Intuition

- They have little semantic content
- There are a lot of them: around 30% of postings for top 30 words

Stop Words

Things with little value

- Common words which would appear to be of little value.
  - e.g. the, a, and, to, be

What is the Intuition

- They have little semantic content
- There are a lot of them: around 30% of postings for top 30 words

However

They are useful in many ways

- You need them for:
  - Phrase queries: “King of Denmark”
  - Various song titles, etc.: “Let it be”, “To be or not to be”
  - “Relational” queries: “flights to London”

Thus, the new trends are

- Good compression techniques means the space for including stop words in a system is very small
- Good query optimization techniques mean you pay little at query time for including stop words.

However

They are useful in many ways

- You need them for:
  - Phrase queries: "King of Denmark"
  - Various song titles, etc.: "Let it be", "To be or not to be"
  - "Relational" queries: "flights to London"

Thus, the new trends are

- Good compression techniques means the space for including stop words in a system is very small
- Good query optimization techniques mean you pay little at query time for including stop words.

Case Folding

Reduce all letters to lower case

- It is a good idea to lower the cases

Longstanding Google example

- – Query C.A.T. – #1 result is for "cats" (well, Lolcats) not Caterpillar Inc

Case Folding

Reduce all letters to lower case

- It is a good idea to lower the cases

Longstanding Google example

- – Query C.A.T. – #1 result is for “cats” (well, Lolcats) not Caterpillar Inc

Further Steps

Stemming and Lemmatization

- This is important to encode the numerical meaning of the document!!!

Outline

Stemming

Definition

- Stemming works by cutting of the end or the beginning of the word.
  - By using common prefixes and suffixes...

Classic Algorithm for English

- The Porter's Algorithm

For Example

- e.g., automate(s), automatic, automation all reduced to automat.

Stemming

Definition

- Stemming works by cutting of the end or the beginning of the word.
  - By using common prefixes and suffixes...

Classic Algorithm for English

- The Porter's Algorithm

For Example

- e.g., automate(s), automatic, automation all reduced to automat.

Stemming

Definition

- Stemming works by cutting of the end or the beginning of the word.
  - By using common prefixes and suffixes...

Classic Algorithm for English

- The Porter's Algorithm

For Example

- e.g., automate(s), automatic, automation all reduced to automat.

For Example

Original Document

- “for example compressed and compression are both accepted as equivalent to compress.”

Into the following

- “for exampl compress and compress ar both accept as equival to compress”

For Example

Original Document

- “for example compressed and compression are both accepted as equivalent to compress.”

Into the following

- “for exampl compress and compress ar both accept as equival to compress”

Outline

Porter's Algorithm

Definition of a Constant

- A consonant in a word is a letter other than A, E, I, O and U, and other than Y preceded by a consonant.

Therefore

- In TOY the consonants are T and Y,
- In SYZYGY they are S, Z, and G.

Therefore

- If a letter is not a consonant it is a vowel

Porter's Algorithm

Definition of a Constant

- A consonant in a word is a letter other than A, E, I, O and U, and other than Y preceded by a consonant.

Therefore

- In TOY the consonants are T and Y,
- In SYZYGY they are S, Z, and G.

Therefore

- If a letter is not a consonant it is a vowel

Porter's Algorithm

Definition of a Constant

- A consonant in a word is a letter other than A, E, I, O and U, and other than Y preceded by a consonant.

Therefore

- In TOY the consonants are T and Y,
- In SYZYGY they are S, Z, and G.

Therefore

- If a letter is not a consonant it is a vowel

Then

Simplicity

- A consonant will be denoted by c, a vowel by v

Thus, we have the following

- A list ccc... of length greater than 0 will be denoted by C,
- A list vvv... of length greater than 0 will be denoted by V

Then

Simplicity

- A consonant will be denoted by c, a vowel by v

Thus, we have the following

- A list **ccc...** of length greater than 0 will be denoted by C,
- A list **vvv...** of length greater than 0 will be denoted by V

Outline

Therefore, we have...

Basic Structures for the Words

CVCV...C
CVCV...V
VCVC...C
VCVC...V

With Final Representation

- $[C](VC)^m[V]$

m is the measure of any word or word part

- The case m = 0 covers the null word.

Therefore, we have...

Basic Structures for the Words

CVCV...C
CVCV...V
VCVC...C
VCVC...V

With Final Representation

- $[C](VC)^m[V]$

m is the measure of any word or word part.

- The case m = 0 covers the null word.

Therefore, we have...

Basic Structures for the Words

CVCV...C
CVCV...V
VCVC...C
VCVC...V

With Final Representation

- $[C](VC)^m[V]$

m is the measure of any word or word part

- The case m = 0 covers the null word.

Examples

For $m = 0$

- TR, EE, TREE, Y, BY

For $m = 1$

- TROUBLE, OATS, TREES, IVY.

For $m = 2$

- TROUBLES, PRIVATE, OATEN, ORRERY.

Examples

For $m = 0$

- TR, EE, TREE, Y, BY

For $m = 1$

- TROUBLE, OATS, TREES, IVY.

For $m = 2$

- TROUBLES, PRIVATE, OATEN, ORRERY.

Examples

For $m = 0$

- TR, EE, TREE, Y, BY

For $m = 1$

- TROUBLE, OATS, TREES, IVY.

For $m = 2$

- TROUBLES, PRIVATE, OATEN, ORRERY.

Outline

Rules

The rules for removing suffix will be of the form

- (condition)$S_1 \rightarrow S_2$

This means that, if a word ends with the suffix S_1

- The stem before S_1 satisfies the given condition S_1 is replaced by S_2.

For Example, (m>1) EMENT $\rightarrow$

- REPLACEMENT to REPLAC

Rules

The rules for removing suffix will be of the form

- (condition)$S_1 \rightarrow S_2$

This means that, if a word ends with the suffix S_1

- The stem before S_1 satisfies the given condition S_1 is replaced by S_2.

For Example, (m>1) EMENT →

- REPLACEMENT to REPLAC

Rules

The rules for removing suffix will be of the form

- (condition)$S_1 \rightarrow S_2$

This means that, if a word ends with the suffix S_1

- The stem before S_1 satisfies the given condition S_1 is replaced by S_2.

For Example, (m>1) EMENT $\rightarrow$

- REPLACEMENT to REPLAC

Other Rules

We have that

- * S – the stem ends with S (and similarly for the other letters).
- * v * – the stem contains a vowel.
- * d – the stem ends with a double consonant (e.g. -TT, -SS).
- * o – the stem ends cvc, where the second c is not W, X or Y (e.g. -WIL, -HOP).

There are many other

- You can take a look at them

Other Rules

We have that

- * S – the stem ends with S (and similarly for the other letters).
- * v * – the stem contains a vowel.
- * d – the stem ends with a double consonant (e.g. -TT, -SS).
- * o – the stem ends cvc, where the second c is not W, X or Y (e.g. -WIL, -HOP).

There are many other

- You can take a look at them

Outline

Recall and Precision

Recall

- The higher the proportion of correct items you require in the selected set (high precision), the fewer of the total correct items you will select (low recall)

Precision

- If you can tolerate a higher proportion of incorrect items in the selected set (low precision), you will capture more of the total correct items (high recall)

Recall and Precision

Recall

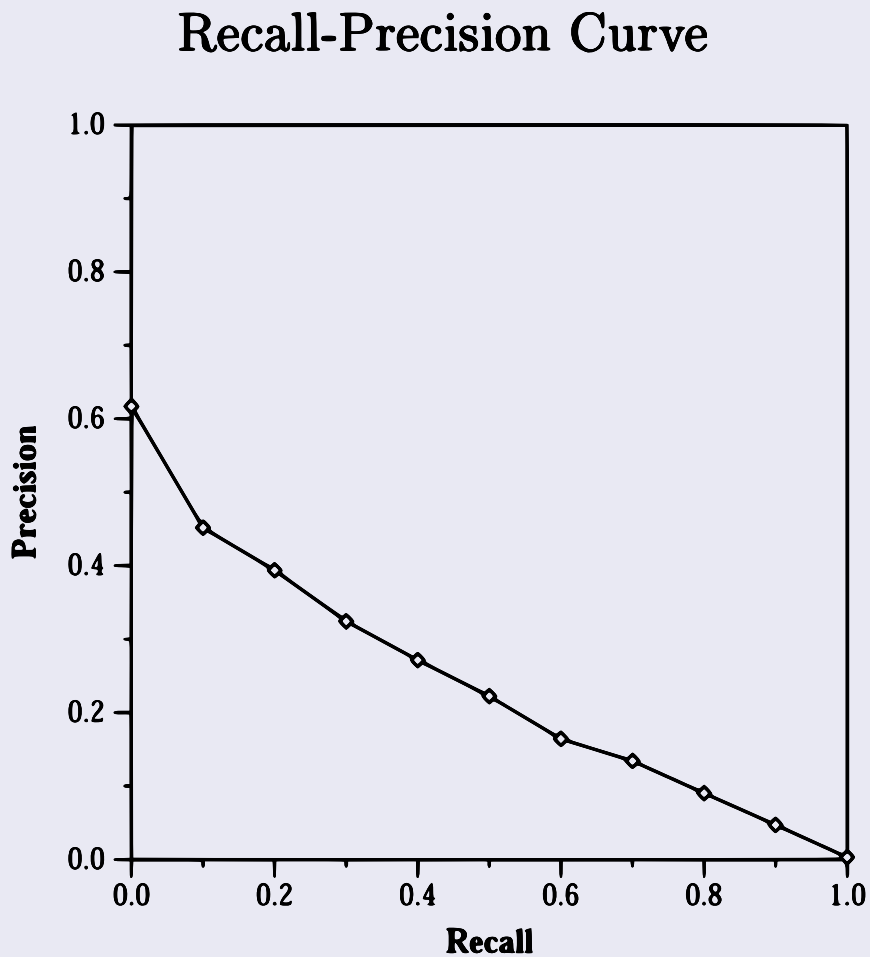
- The higher the proportion of correct items you require in the selected set (high precision), the fewer of the total correct items you will select (low recall)

Precision

- If you can tolerate a higher proportion of incorrect items in the selected set (low precision), you will capture more of the total correct items (high recall)

Using Stemmed Documents

And a Sequence of Stemmed Quereies



Outline

What is Lemmatization?

On the other hand

- it takes into consideration the morphological analysis of the words.

The Algorithm needs

- to have detailed dictionaries

What is Lemmatization?

On the other hand

- it takes into consideration the morphological analysis of the words.

The Algorithm needs

- to have detailed dictionaries

For examples

STUDIES

Morphological Information	Lemma
Third Person, Singular Number Present Tense Study	Study

SUDYING

Morphological Information	Lemma
Gerund form of the verb study	Study

Properties

For examples

STUDIES

Morphological Information	Lemma
Third Person, Singular Number Present Tense Study	Study

SUDYING

Morphological Information	Lemma
Gerund form of the verb study	Study

Properties

For examples

STUDIES

Morphological Information	Lemma
Third Person, Singular Number Present Tense Study	Study

SUDYING

Morphological Information	Lemma
Gerund form of the verb study	Study

Properties

Outline

Rule Based

Here, a system of rules

- It is used to find the root of the words

Therefore

- As in Porter's Algorithm, it takes time to be developed...

An it needs

- To have a dictionary

Rule Based

Here, a system of rules

- It is used to find the root of the words

Therefore

- As in Porter's Algorithm, it takes time to be developed...

An it needs

- To have a dictionary

Rule Based

Here, a system of rules

- It is used to find the root of the words

Therefore

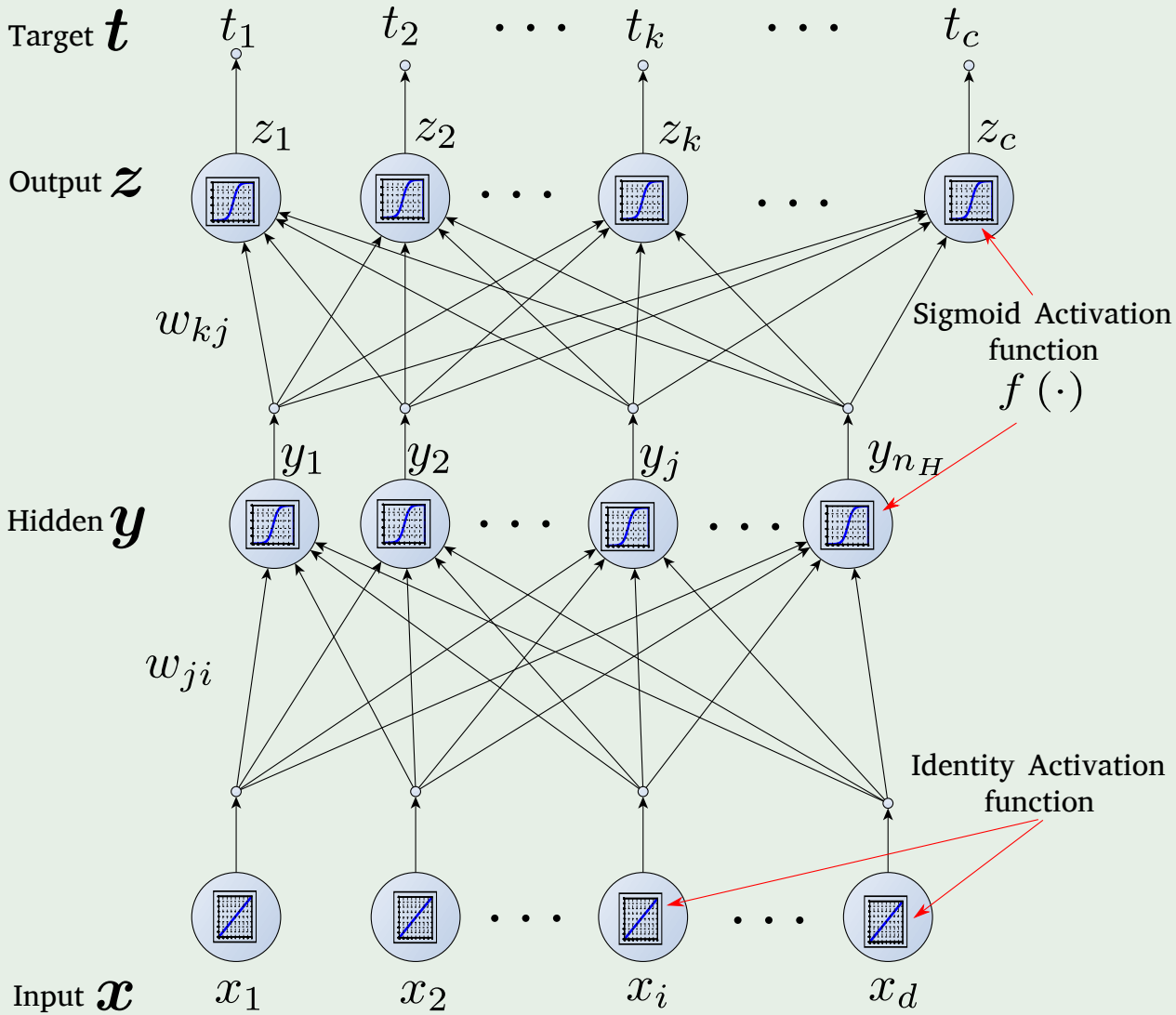
- As in Porter's Algorithm, it takes time to be developed...

An it needs

- To have a dictionary

Neural Networks

They are using neural networks right now...



However

It requires to have labeled data

- Again takes time...